Recitation 3

Regularization: Motivation and Effect

Colin

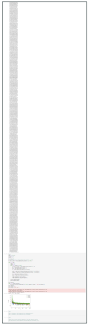
CDS

Feb. 09, 2022

Logistics

- Submitting HWs
    - Ensure they are eligible
        - We encourage you to use LaTeX (Useful skill to learn)
    - Ensure they are in the correct orientation
    - Do not print out irrelevant information
    - Do not cite materials to support your proof
- Late submissions
    - Late days

Examples



Motivation for learning the math/proof

- We are not SDE or BA.
  - Your task is to make a decision through modeling.
  - Understand how each model behaves.
  - Everything will lie to you, not math
- Be able to fix/alter/adjust your model when it fails.
  - Knowing what is expected to happen, what is not.
  - Anyone can copy model from github.
    - Few can diagnose when it doesn't work
  - Help you to build up intuition.

Regularization and its effects

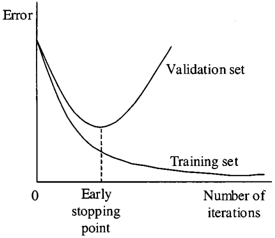
Motivation

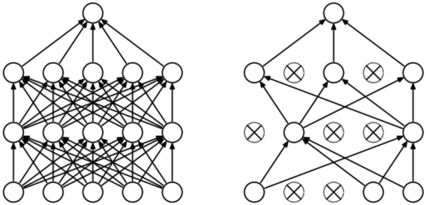
- Hard to choose a good hypothesis space.
  - Knowing too little about the data/truth
- If the space is too small
  - Cannot model the data accurately
- If the space is too large
  - Overfit the training data
  - Amazing in training; Useless when deployed
- Solution:
  - Start with a large space, then shrink it down

Types of Regularization

- Implicit Regularization
  - Initialization
  - Training strategy
  - Model structure
- Explicit Regularization (what we refer to in this course)
  - Classics (what we will discuss today)
    - L1 & L2 & Elastic Net
  - Others
    - Early stopping
    - Data augmentation
    - Dropouts

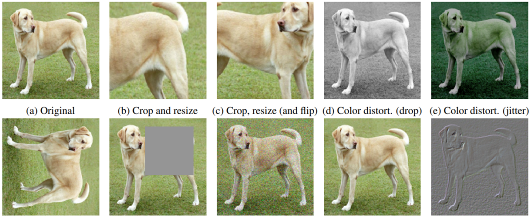
Examples of other types of regularization





(a) Standard Neural Net

(b) After applying dropout.



(a) Original (b) Crop and resize (c) Crop, resize (and flip) (d) Color distort. (drop) (e) Color distort. (jitter)

(f) Rotate {90°, 180°, 270°} (g) Cutout (h) Gaussian noise (i) Gaussian blur (j) Sobel filtering

L2 (Ridge) and L1(Lasso) Regularization

L2 (Ridge)

$$\hat{w} = \underset{w \in R^d}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left\{ w^T x_i - y_i \right\}^2 + \lambda \|w\|_2$$

L1 (Lasso)

$$\hat{w} = \underset{w \in R^d}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left\{ w^T x_i - y_i \right\}^2 + \lambda \|w\|_1$$

Effect on linearly dependent features

- Suppose we have one feature x_1.
- Response variable y.
- The ERM is

$$\hat{f}(x_1) = 4x_1$$

- What happens if we get a new feature x2,
  - but $x_2 = x_1$?

Effect on linearly dependent features

- New feature x2 gives no new information.
- ERM is still

$$\hat{f}(x_1) = 4x_1$$

- Now there are some more ERMs:

$$\hat{f}(x_1, x_2) = 2x_1 + 2x_2$$
$$\hat{f}(x_1, x)2) = x_1 + 3x_2$$
$$\hat{f}(x_1, x_2) = 8x_1 - 4x_2$$

- What if we introduce L1 or L2 regularization?

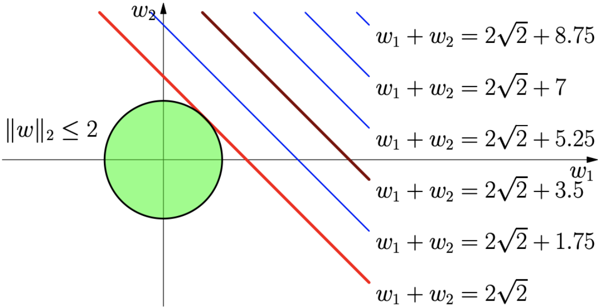
Effect on linearly dependent features

- $f(x_1, x_2) = w_1 x_1 + w_2 x_2$ is an ERM iff $w_1 + w_2 = 4$.
- Consider the L1 and L2 norms of various solutions:

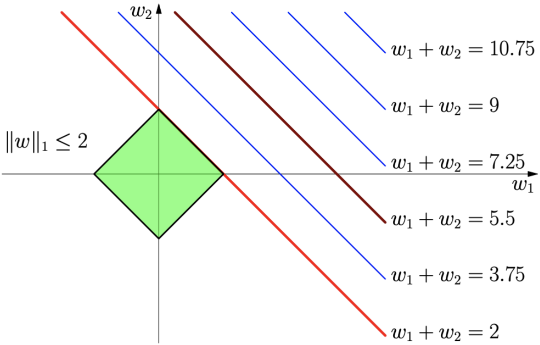
w_1	w_2	$\sum \lvert w_i \rvert_1$	$\sum \lvert w_i \rvert_2^2$
4	0	4	16
2	2	4	8
1	3	4	10
8	-4	12	80

- $|w|_1$ doesn't discriminate, as long as all have same sign
- $|w|_2$ minimized when weight is spread equally

L2 Contour Line



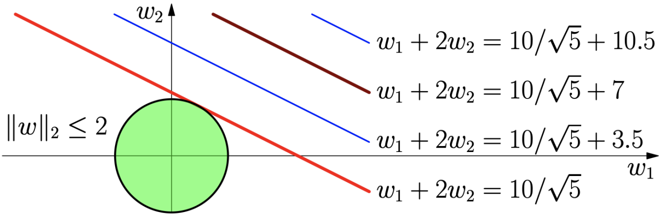
L1 Contour Line



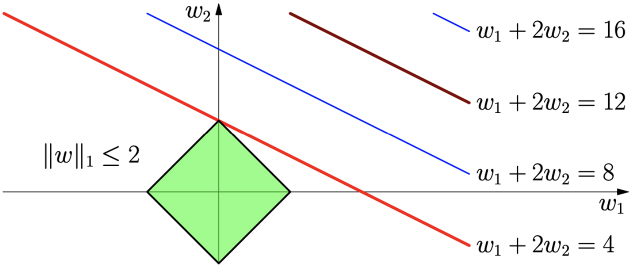
Effect on linearly dependent features

- Now lets consider the case where $x_2 = 2x_1$
- Then any model satisfies $2w_2 + w_1 = 4$ will be an ERM.
  - Suppose we are still dealing with the previous setup
    - $\hat{f}(x_1, x_2) = 2x_1 + x_2$
    - $\hat{f}(x_1, x_2) = 3x_2 + 0.5x_2$
    - $\hat{f}(x_1, x_2) = 6x_2 - x_2$
- How would the regularization change the outcome?

L2 Contour Line



L1 Contour Line



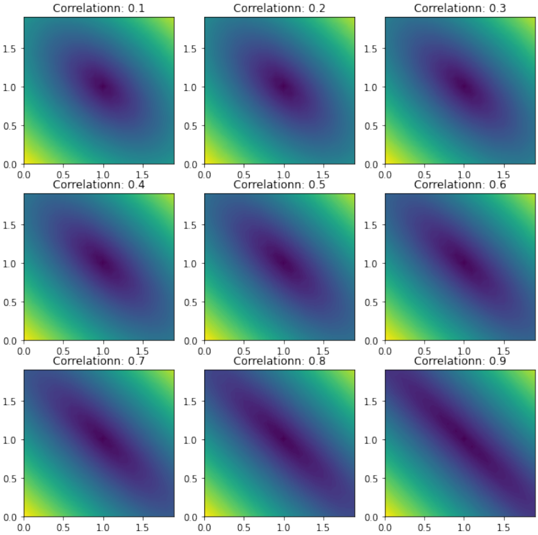
Summary

- For identical features
  - L1 regularization spreads weight arbitrarily (all weights same sign)
  - L2 regularization spreads weight evenly
- Linearly related features
  - L1 regularization chooses variable with larger scale, 0 weight to others
  - L2 prefers variables with larger scale – spreads weight proportional to scale

Note on contour lines

- Recall our discussion of linear predictors $f(x) = w^T x$ and square loss.
- Sets of w giving same empirical risk (i.e. level sets) formed ellipsoids around the ERM.
- With x_1 and x_2 linearly related, $X^T X$ has a 0 eigenvalue.
- So the level set $\{ \hat{w} \mid (w - \hat{w})^T X^T X (w - \hat{w}) = c \}$ is no longer an ellipsoid.
- It's a degenerate ellipsoid - that's why level sets were pairs of lines in this case

Note on contour lines

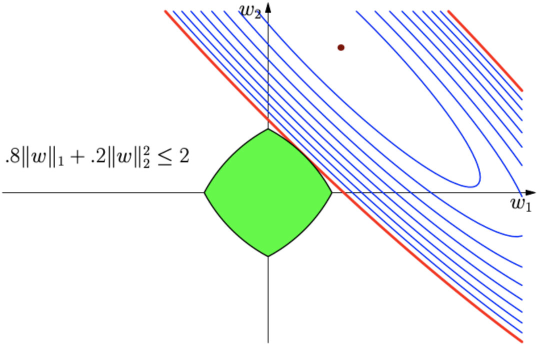


Elastic Net

A way of combining $L1$ and $L2$ regularization:

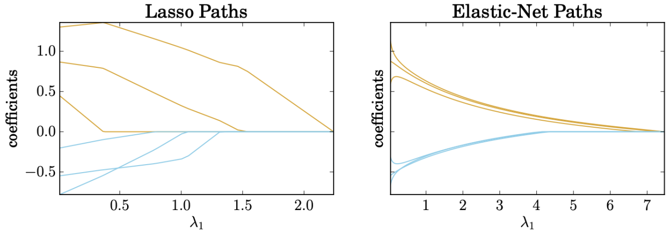
$$\hat{w} = \arg\min_{w \in d} \frac{1}{n} \sum_{i=1}^{n} \left\{ w^T x_i - y_i \right\}^2 + \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2^2$$

Elastic Net



A not so inspiring way of compromising.

Elastic Net



- Ratio of $L2$ to $L1$ regularization roughly $2:1$.

Generalization into more complicated models

- The goal is to make model remember only the **relevant** information.
- Reduce the model's dependency of each feature as much as possible.
    - Methods may vary when we have billions of parameters.